Real Time Crowd Detection And Counting Using Switching Convolutional Neural Networks

R. Prabhu, M.E., Arun Balaji A, Balachandran E

***Abstract*— The concept of crowd counting is to count the number of people in a locality using a live camera. Here we go with the real time crowd detection of the streaming video, as we can monitor the number of people visiting a particular place for security purpose, surveillance of sports event, political rallies, stampedes and various applications. The factors that reduce the accuracy on crowd counting are similar appearance of people, improper projection of head from the view point of the camera. In this work, switching convolutional neural networks (S-CNN) is used for increasing the accuracy of crowd detection and counting. In earlier approaches the inter scene variation is not considered but while using S-CNN the inter scene variation and semantic analysis is considered to improve the estimation of the count. The Shanghai-Tech data set is used to train and evaluate the CNN model. S-CNN model is an algorithm used to segregate the patches of the frames of the video based on the density of the crowd and to count with more accuracy. The training data set includes 300 images of various places and test images are taken from the frames of the videos.**

***Index Terms*— Image processing, Convolutional Neural Network, Switching Convolutional Neural Network, Crowd counting, Surveillance**

I. INTRODUCTION

Crowd analysis has important applications in the scenario of political rallies, Sport events, Surveillance, etc. Over the last few years, researchers have attempted to address the issue of crowd counting can be done by the approaches such as detection-based counting, clustering-based counting and regression-based counting. The initial work on regression-based methods mainly used in handcrafted features and the more recent works use Convolutional Neural Network (CNN) based approaches. The CNN-based approaches have some important improvements over previous feature based methods, thus, motivating more researchers to explore CNN-based approaches further for related crowd analysis problems. The CNN has contains the two important layers are convolutional layer and pooling layer. In our method, we used the patch based inference approach. There may be varying crowd density with in a single image that results in different pixel sizes per head at different positions in the image. So in order to overcome this issue we have used three CNN regressors each suitable for different crowd density.so the non-overlapping patches are generated from the frame which is allocated to the regressor based on their density by switching network. The switching network is trained in order to get correct estimation. Finally the density map generated for each patch by the regressor is combined and the count is calculated. Then after counting of the first frame, the next frame can be taken for counting likewise its counting and detecting the crowd of the real time video.

Prabhu R, M.E., Assistant Professor, Department Of Electronics And Communication Engineering, St.Joseph's College Of Engineering, India
Arun Balaji A, Student, Department Of Electronics And Communication Engineering, St.Joseph's College Of Engineering, India
Balachandran E, Student, Department Of Electronics And Communication Engineering, St.Joseph's College Of Engineering, India

II. RELATED WORKS

We introduce the various methods which are related to our work. In yearly crowd detection is based on face detection using viola-jones algorithm [4]. But this algorithm is not efficient as much required because some images may have faces with improper positions like they may get overlapped, may be people face beyond one another face and it may be captured at wrong position. After this view point invariant approach is introduced. It is based on the orientation of the camera, but it also not as much as accurate [1]. Because some position of cameras does not support with this view point invariant approach method. Later the counting is done with help of human position based templates [5]. In this method the human shaped templates are compared with the human shapes in the images, but it also not giving that much efficient as expected. Then the generic head detector is applied to get better output [7]. Followed by this Convolutional Neural Network is a introduced to train the system with multiple datasets, which will increase the accuracy. Then the crowd counting from the video can be done by image from the framing cluster. In Multi-column convolutional neural network, the crowd detection can be done by the density map for the images [11]. In our approach, we use the adaptive kernel algorithm for the density map. We counting the number of people by adding the number of one's in the density map generated matrix.

III. PROPOSED METHOD

In our method we crowd counting from the video can be done by using switching convolutional neural networks.

Figure 1.a

Figure 1.b

Figure 1,a and Figure 1.b shows the frames of the video which will be used for further processing.

The frame is split into nine non over lapping image patches with each image 1/3rd of the size of the original image. Then the patch is allocated to individual CNN regressor depending on the crowd intensity in that particular patch. For this we consider switching CNN architecture (Switch-CNN) that relays patches from a grid within a crowd scene to independent CNN regressors with help of a switch classifier

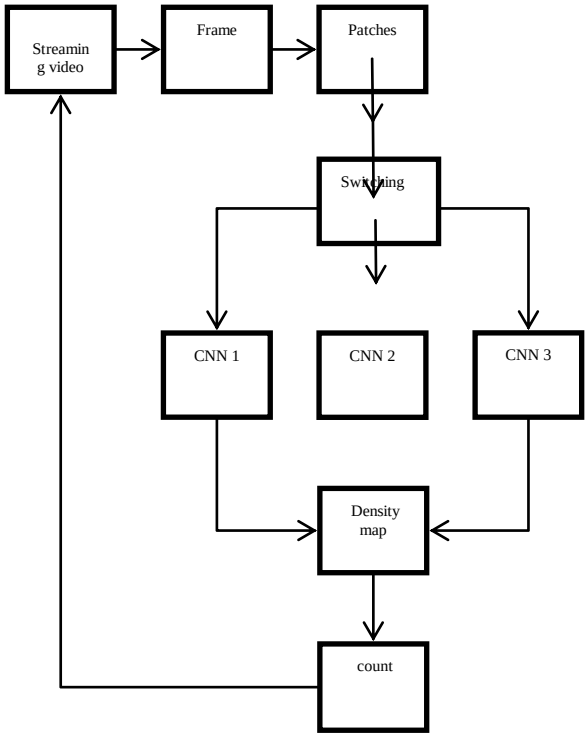

Figure 2: Flow diagram of the methodology

. The independent CNN regressors are chosen with different receptive fields and field-of-view as in multi-column CNN networks to augment the ability to model large scale variations. A particular CNN regressors are trained on a crowd scene patch if the performance of the regressor on the patch is the best. A switch classifier is trained alternately with the training of multiple CNN regressors to correctly relay a patch to a particular regressor. The salient properties that make this model excellent for crowd analysis are the ability to model large scale variations the facility to leverage local variations in density within a crowd scene. The ability to leverage local variations in density is important as the weighted averaging technique used in multi-column networks to fuse the features is global in nature. Once all patches of a particular image is allocated to CNN regressors the density map of each patch will be generated and from the density map the crowd count of each patch is calculated and their sum is added together to get the total head count. These all the operation are done within the seconds. Then after that the next frame of the streaming video is taken and its follow the similar steps mentioned above.

A. Training

In our method, we used differential training and coupled training. The differential training on the switch helps to choosing the best regressor for a crowd intensity of the frame. The three CNN regressor has a different filter size that will helps to detecting the blob based on the density of crowd in the patch. The process of alternating the switch training and switched training of the CNN regressors is repeated until the correct regressor is chosen. After the coupled training the best CNN is considered. For training we use the Shanghai tech dataset (part A), it consists of 382 training images and their ground truth value. In SCNN, adaptive kernel algorithm is to generate the ground truth density for the frame. The ground truth density will help to predict the crowd value with greater accuracy.

B. Performance analysis

The performance of the proposed system can be evaluated by Mean Absolute Error (MAE). The Mean Average Error is defined as the difference between the ground truth and the count value of our proposed system.

$$\mathbf{MAE} = \frac{1}{N}\sum_{i=1}^{N} |C_i - C_i^{GT}|$$

IV. RESULTS AND DISCUSSION

The frame from the video is divided into 9 non overlapping patches. The patch can be switched to the correct regressor based on the intensity of crowd in the patch. Then the regressor will generate matrix file for the corresponding patch. The matrix file will contain 0s and 1s in its cells. The 1s represent the location of the head in the patch and the rest of the matrix is filled with 0s.The number of 1's in the mat file of the patch represent the count and finally the count of all patches are added to get the total count of the image.

Figure 3: Patches of the frame

The 9 non overlapping patches are separated from the frame is shown in Figure 2.

Figure 4.a.

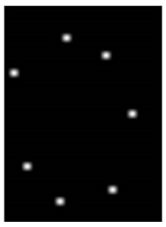

Figure 4.b.

Figure 5.a.

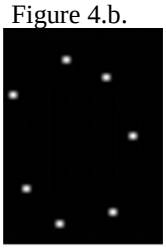

Figure 5.b.


Figure 6.a.

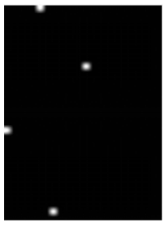
Figure 6.b.


Figure 7.a.

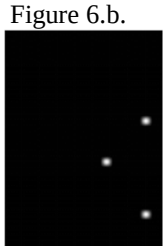
Figure 7.b.


Figure 8.a.

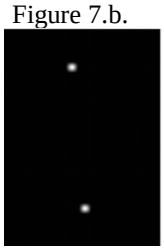
Figure 8.b.


Figure 9.a.

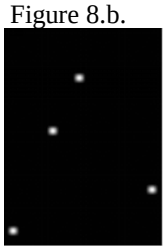
Figure 9.b.


Figure 10.a.

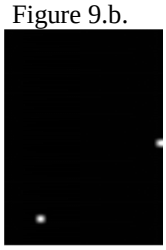
Figure 10.b.


Figure 11.a.

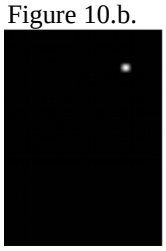
Figure 11.b.


Figure 12.a.

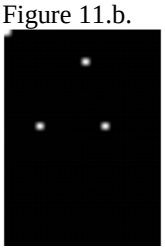
Figure 12.b.

Figure 4 to Figure 12 shows the patches of the frame and their respective density mapping.

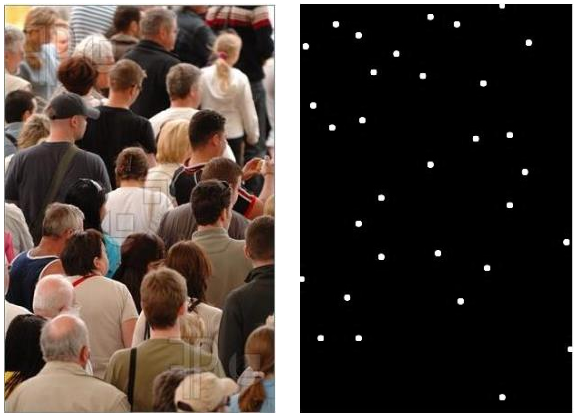
Figure 13: Original image and the combined density mapping

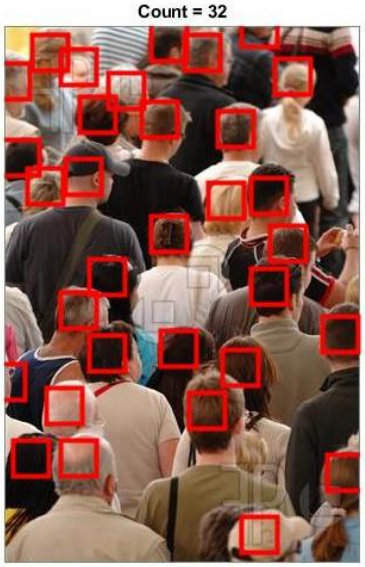

Figure 14: Detected crown marked with help of density mappings

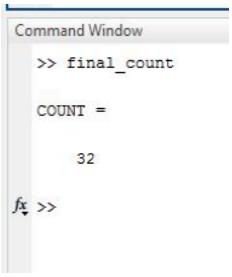

Figure 15

Figure 15 shows the final count of the frame.

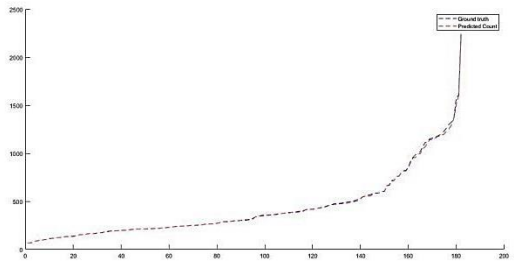
Figure 16: Comparison graph between ground truth and predicted value

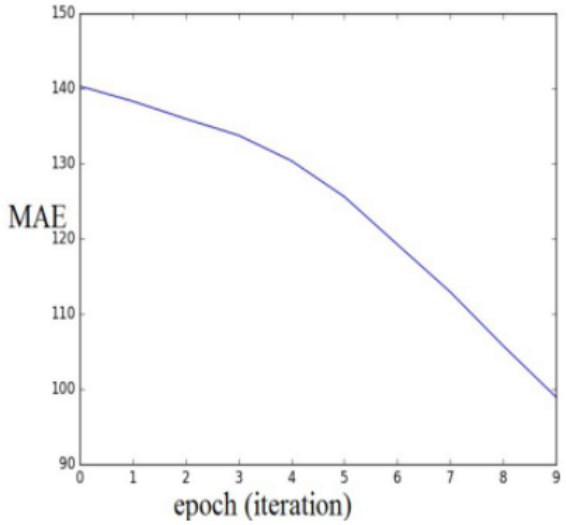

Figure 17: Mean Average Error

Figure 17 shows the MAE which will be decreased rapidly when the number of iteration of training is increased. More iteration will reduce the MAE which will reflect in accuracy.

V. Conclusion

In this project, crowd detection is performed using convolutional neural networks. The advantages of switching convolutional neural network that leverages intra-image crowd density variation to improve the accuracy of localization of the people. Based on the training of the CNN regressor we got a count of the number of peoples from the images with good accuracy. The proposed algorithm is implemented using Shanghai Tech part A dataset and observed that it produces the Mean Average Error value is 98.87. When the number of iteration is increased the MAE value is further reduced. In future , we implement the same algorithm to the multiple datasets like shanghai dataset Part B, The UCF_CC_50 dataset, The UCSD dataset, etc., that will help to increases the accuracy of counting.

REFERENCES

[1] Kong D, Gray D, and Tao H, "A viewpoint invariant approach for crowd counting," in Proc. IEEE ICPR, vol. 3, 2006, pp. 1187–1190.
[2] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 7, pp. 1198–1211, 2008.
[3] Ryan D, Denman S, Fookes C, and Sridharan S, "Crowd counting using multiple local features," in Proc. IEEE DICTA, 2009, pp. 81–88.
[4] Theo Ephraim, Tristan Himmelman and Kaleem Siddiqi, "Real time viola-jones face detection ",in Proc.IEEE Conf.2009.
[5] Z. Lin and L. S. Davis, "Shape-based human detection and segmentation via hierarchical part-template matching," IEEE Trans. Pattern Anal.Mach. Intell., vol. 32, no. 4, pp. 604–618, 2010.
[6] Y. Bo and C. C. Fowlkes, "Shape-based pedestrian parsing," in Proc.IEEE Conf. CVPR, 2011, pp. 2265–2272.
[7] V. B. Subburaman, A. Descamps, and C. Carincotte, "Counting peoplein the crowd using a generic head detector," in Proc. IEEE Conf. AVSS, 2012, pp. 470–475.
[8] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in Proc. IEEE Conf. CVPR, 2013, pp. 2547–2554.
[9] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in Proc. ECCV. Springer, 2014, pp. 818–833.
[10] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in Proc. IEEE Conf. CVPR,2015, pp. 833–841.
[11] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in Proc. IEEE Conf. CVPR, 2016, pp. 589–597

Prabhu R, M.E., Assistant Professor, Department Of Electronics And Communication Engineering, St.Joseph's College Of Engineering, India
Arun Balaji A, Student, Department Of Electronics And Communication Engineering, St.Joseph's College Of Engineering, India
Balachandran E, Student, Department Of Electronics And Communication Engineering, St.Joseph's College Of Engineering, India